ApacheCon

Best of breed
httpd, forrest, solr and droids

Thorsten Scherler
Sociedad Andaluza para el Desarrollo
de la Sociedad de la Información S.A.U. (SADESI)

thorsten@apache.org
ApacheCon EU 2009, Amsterdam, 26 March 2009

ApacheCon

Leading the Wave of Open Source

agenda

- use case – official gazette of the Junta de Andalucia
- architecture – the big picture
- forrest – generate static page
- solr – enable search
- droids – task automation
- httpd – answer high traffic requests
- next steps

use case

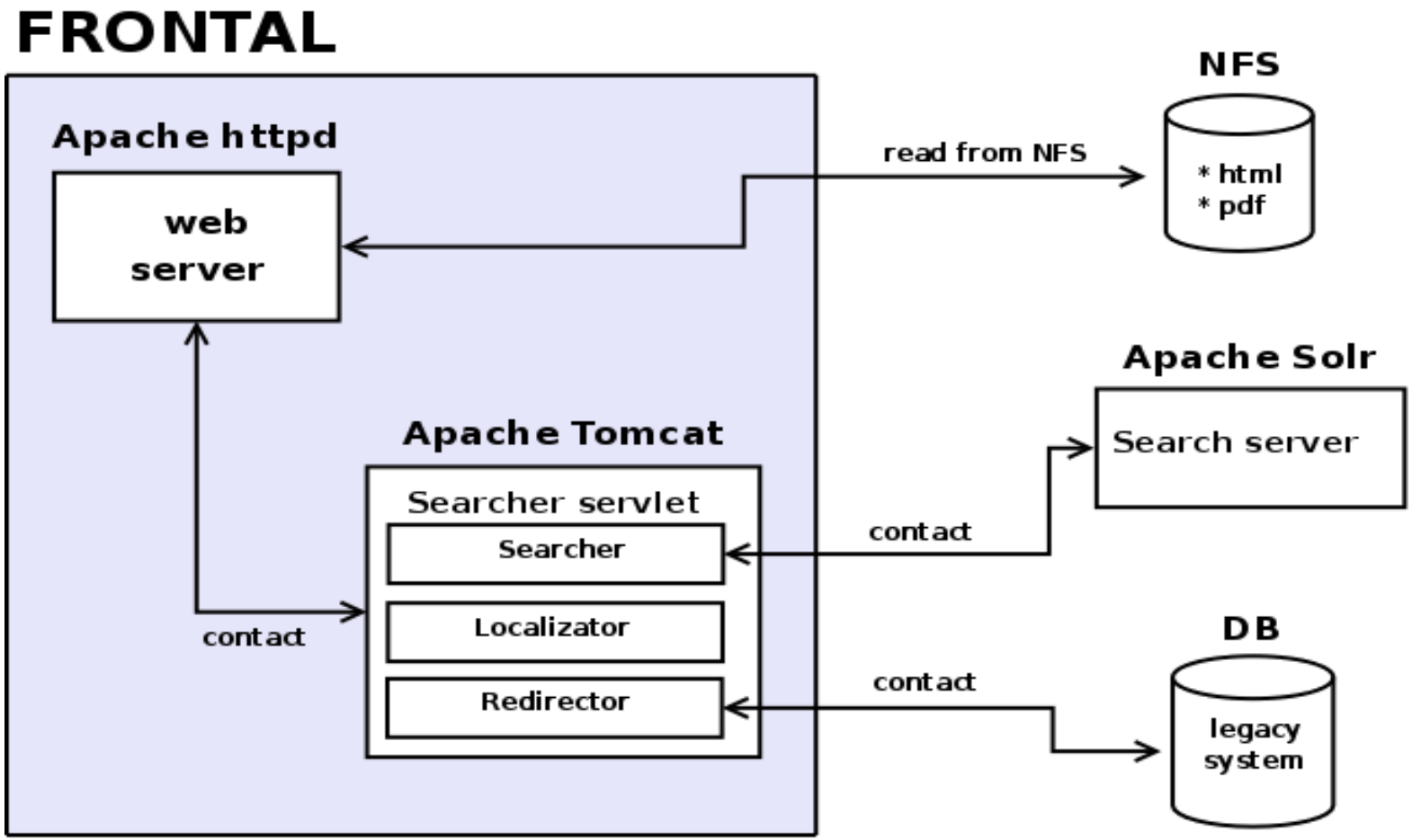
- http://www.juntadeandalucia.es/boja
- high traffic site
- high quantity of static content

- statistic 2008-08
  - site views: 1.5 million (60%)
  - pages (2.66/view): 4 million (60%)
  - requests: 22 million (35%)
  - upload: 300 GB (72%)

use case

- http://www.juntadeandalucia.es
- portal statistic 2008-08
  - site views: 2 million
  - pages (3.22/view): 6.7 million
  - requests: 62 million
  - upload: 420 GB

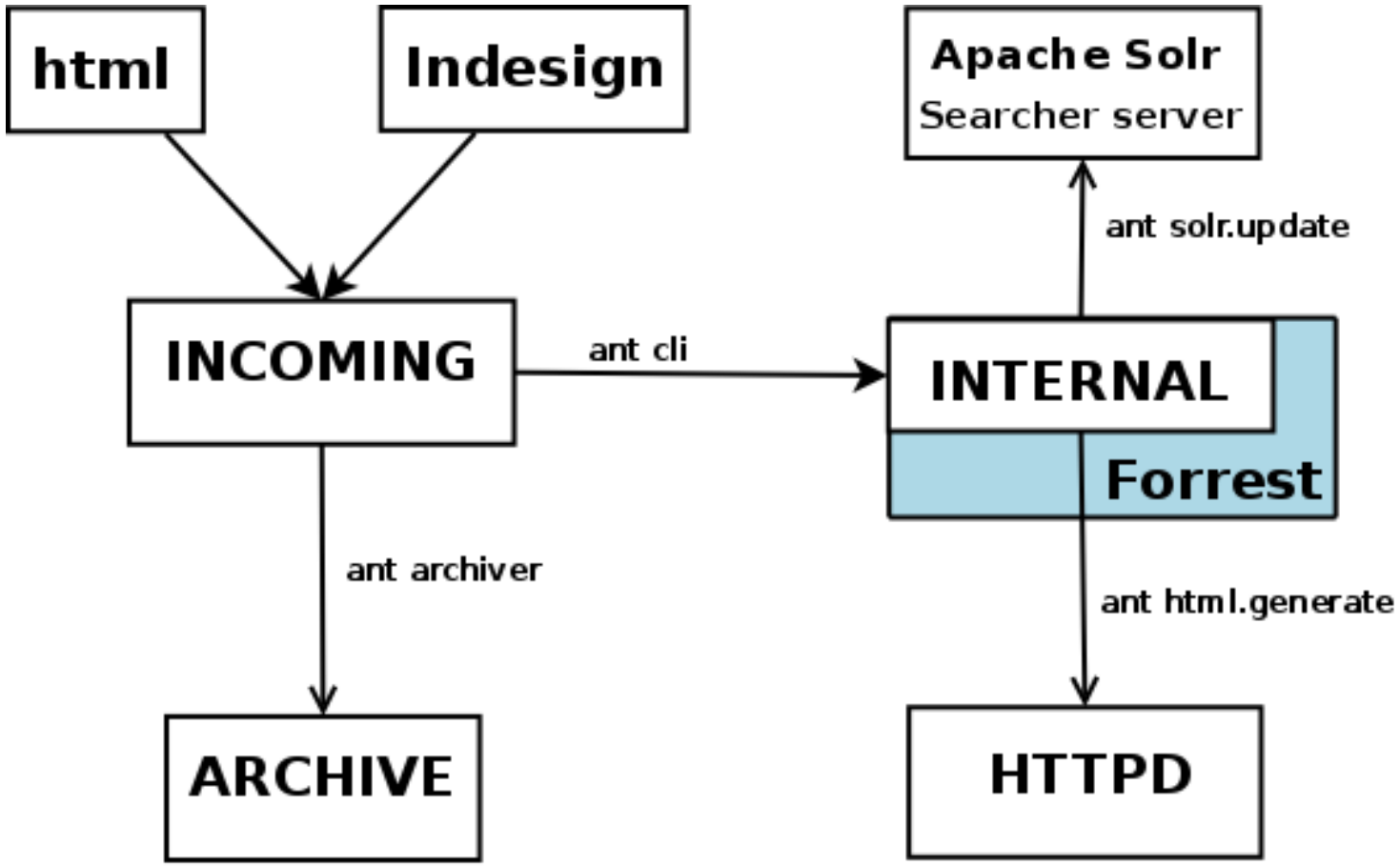
front-end architecture



back-end

- daily updates of latest official gazette
  - input formats (pdf, xml & html)
  - html pages (gazette + global)
  - various pdf generation (fascicle, ...)
  - indexing of content (per disposition)

back-end architecture



forrest

Apache Forrest is a publishing framework that transforms input from various sources into a unified presentation in one or more output formats.

Forrest can generate static documents, or be used as a dynamic server, or be deployed by its automated facility.

forrest

- automated facility for page generation

```
<!-- macro for calling forrest site (with urifile)-->
<macrodef name="site-set">
  <attribute name="uri" />
  <attribute name="build" />
  <attribute name="urifile" />
  <attribute name="followLinks" />
  <sequential>
    <antcall target="site">
      <param name="project.home" location="${exporter.home}" />
      <param name="project.start-uri" location="@{uri}" />
      <param name="project.build-dir" location="@{build}" />
      <param name="project.urifile" location="@{urifile}" />
      <param name="project.followLinks" value="@{followLinks}" />
    </antcall>
  </sequential>
</macrodef>
```

forrest

- The aim of the dispatcher concept is to provide a flexible framework for creating site specific layout in different formats.
  - hook's are containers that are used for layout reasons.
  - contract's are functionality or extra content that a theme can use to display the request.

forrest

- structurer to design the pages

```
<forrest:structure type="html" hooksXpath="/html/body">
  <!-- ... -->
  <forrest:hook id="barra_lateral_izq">
    <forrest:contract name="nav-boja-servicios"/>
    <jx:if test="${!isCalendar}">
      <forrest:contract name="content-pdf-link"
        dataURI="cocoon://${niveles[0]}/${niveles[1]}/${niveles[2]}/hasPdf.xml">
        <forrest:property name="number" value="${niveles[2]}"/>
        <forrest:property name="year" value="${niveles[1]}"/>
      </forrest:contract>
    </jx:if>
  </forrest:hook>
  <!-- ... -->
</forrest:structure>
```

forrest

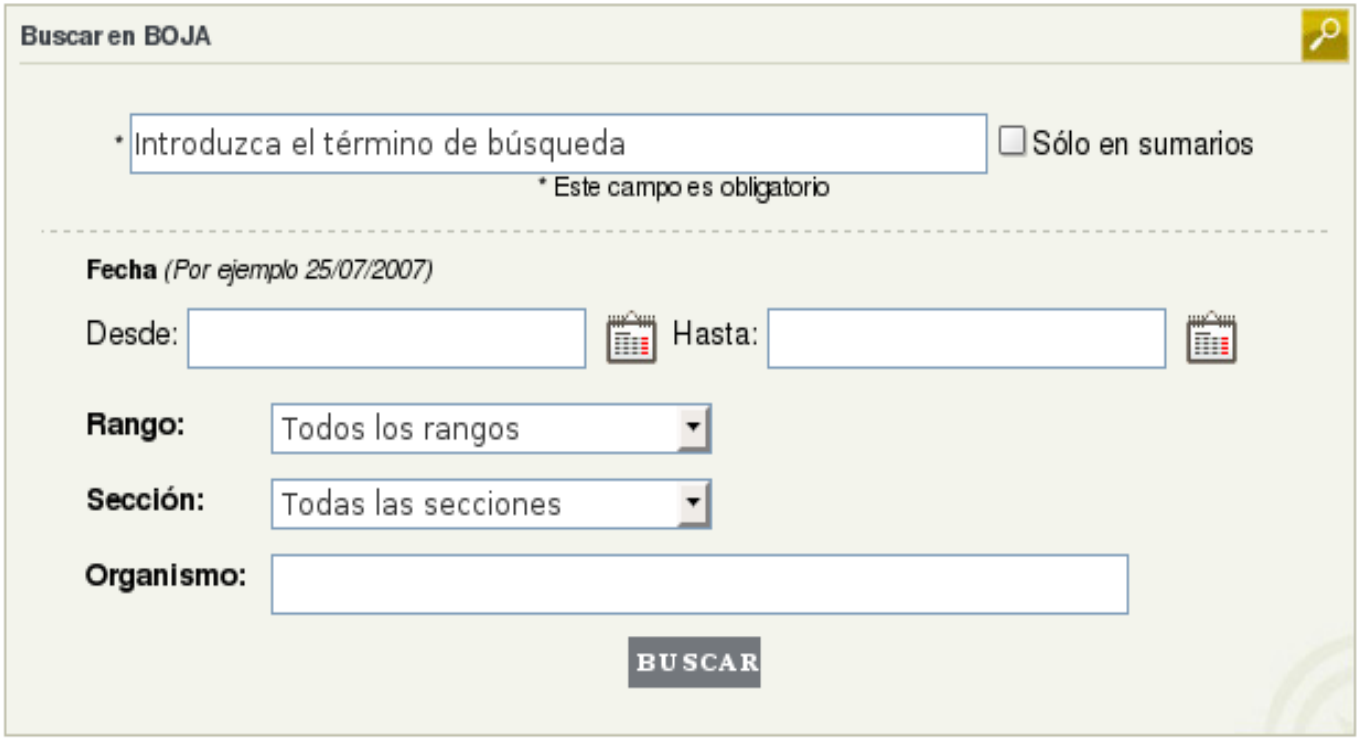
- Forrest solr plugin generates solr documents from xdos.
  - When run with the dispatcher allows you to update solr with the content of your site while generating it (solr-add contract).
  - In dynamic mode it provides a GUI to manage your project in solr (solr-actionbar contract) and a search interface (solr-search contract) to search your solr server.

solr

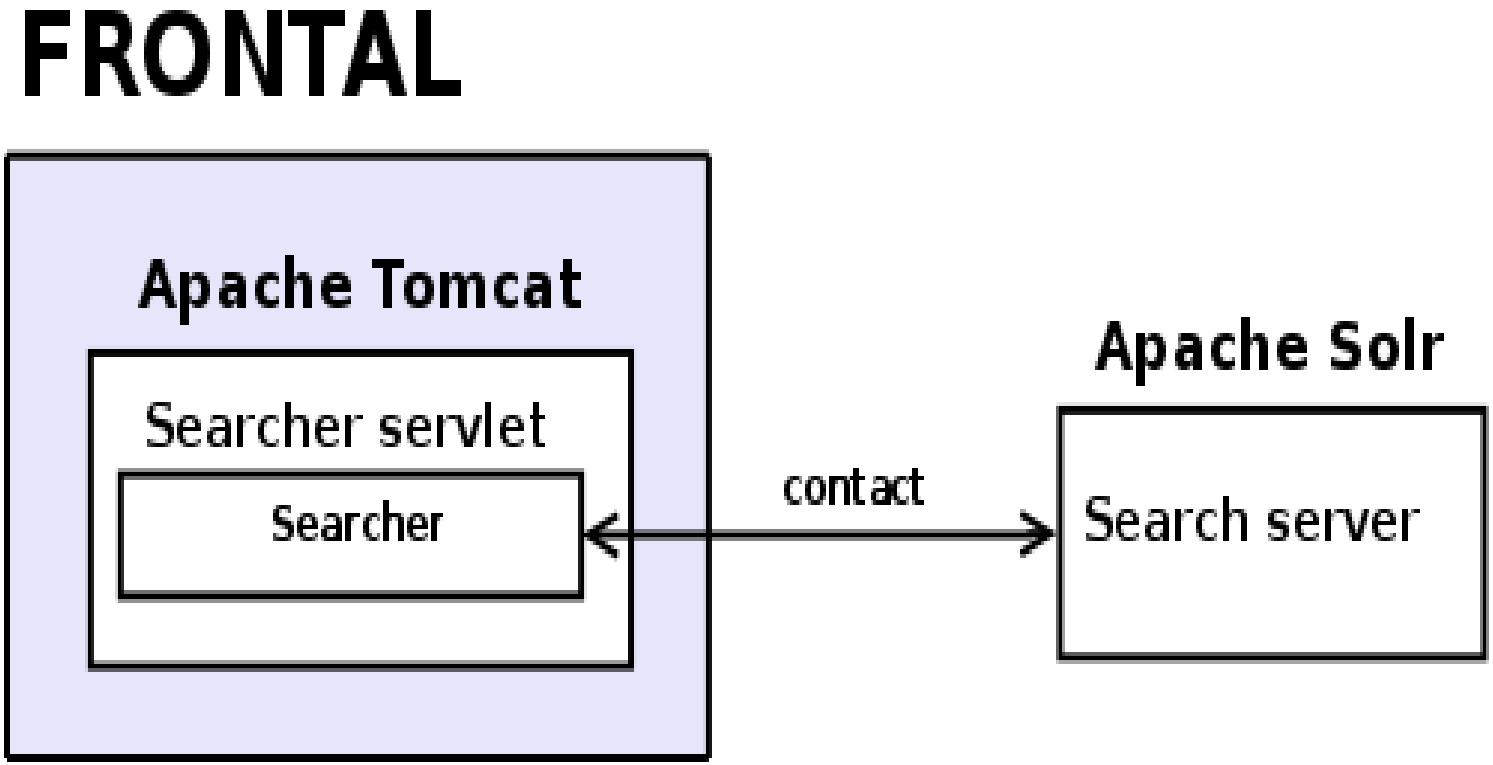
Solr is an open source enterprise search server based on the Lucene Java search library, with XML/HTTP and JSON APIs, hit highlighting, faceted search, caching, replication, a web administration interface and many more features.

solr

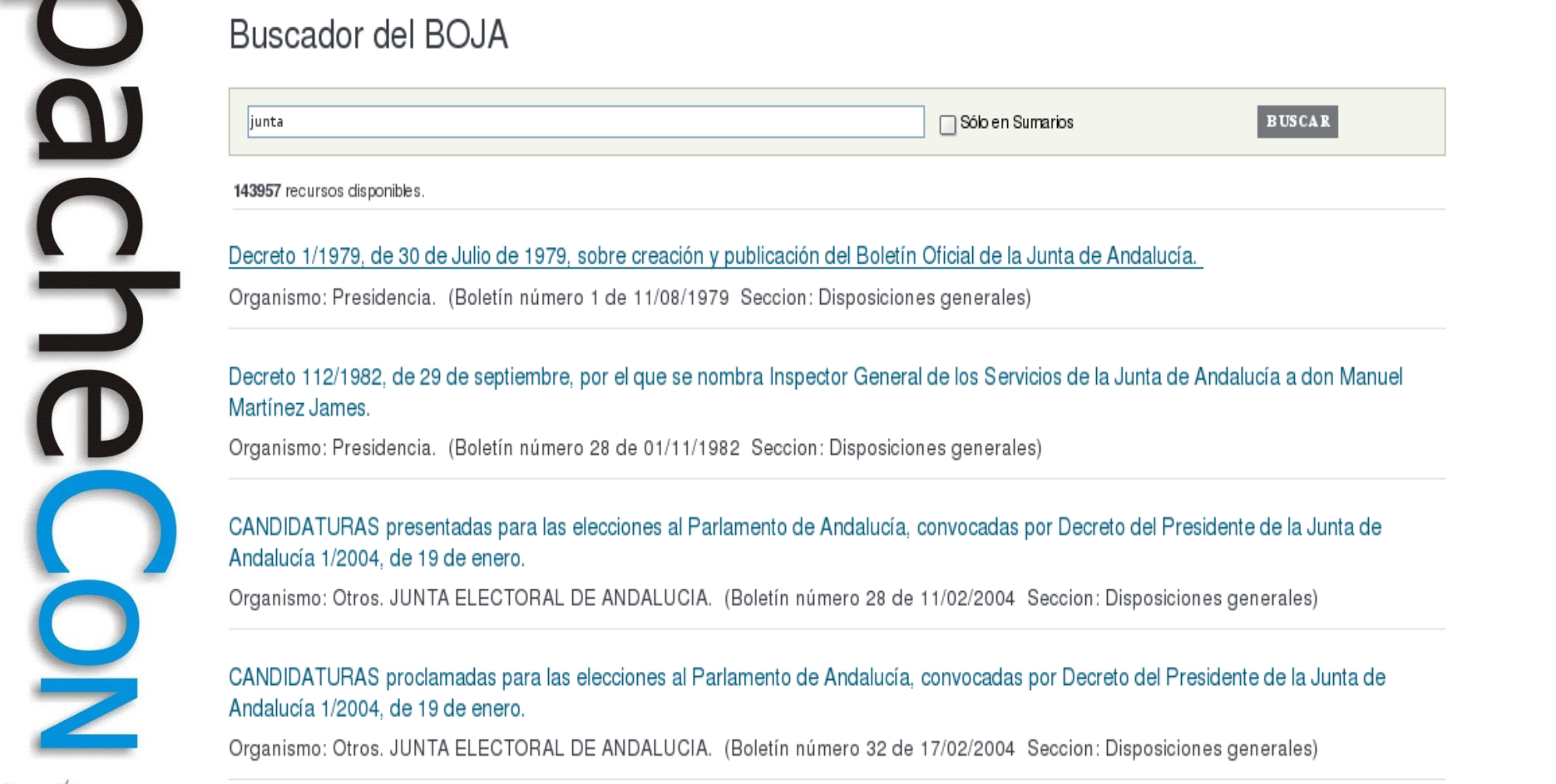
Buscador del BOJA



solr


FRONTAL
Apache Tomcat
Searcher servlet
Searcher
contact
Apache Solr
Search server

solr


Buscador del BOJA
junta
Sólo en Sumarios
BUSCAR
143957 recursos disponibles.
Decreto 1/1979, de 30 de Julio de 1979, sobre creación y publicación del Boletín Oficial de la Junta de Andalucía.
Organismo: Presidencia. (Boletín número 1 de 11/08/1979 Seccion: Disposiciones generales)
Decreto 112/1982, de 29 de septiembre, por el que se nombra Inspector General de los Servicios de la Junta de Andalucía a don Manuel Martínez James.
Organismo: Presidencia. (Boletín número 28 de 01/11/1982 Seccion: Disposiciones generales)
CANDIDATURAS presentadas para las elecciones al Parlamento de Andalucía, convocadas por Decreto del Presidente de la Junta de Andalucía 1/2004, de 19 de enero.
Organismo: Otros. JUNTA ELECTORAL DE ANDALUCIA. (Boletín número 28 de 11/02/2004 Seccion: Disposiciones generales)
CANDIDATURAS proclamadas para las elecciones al Parlamento de Andalucía, convocadas por Decreto del Presidente de la Junta de Andalucía 1/2004, de 19 de enero.
Organismo: Otros. JUNTA ELECTORAL DE ANDALUCIA. (Boletín número 32 de 17/02/2004 Seccion: Disposiciones generales)

use case

- statistic 2008-08
  - searches: 10.000/daily
  - site index: 1.7 GB
  - numDocs : 314.348

droids

Droids is an intelligent standalone robot framework that allows to create and extend existing droids (robots).

A droid can automatically seek out relevant online information based on the user's specifications and invoke custom handler on this information.

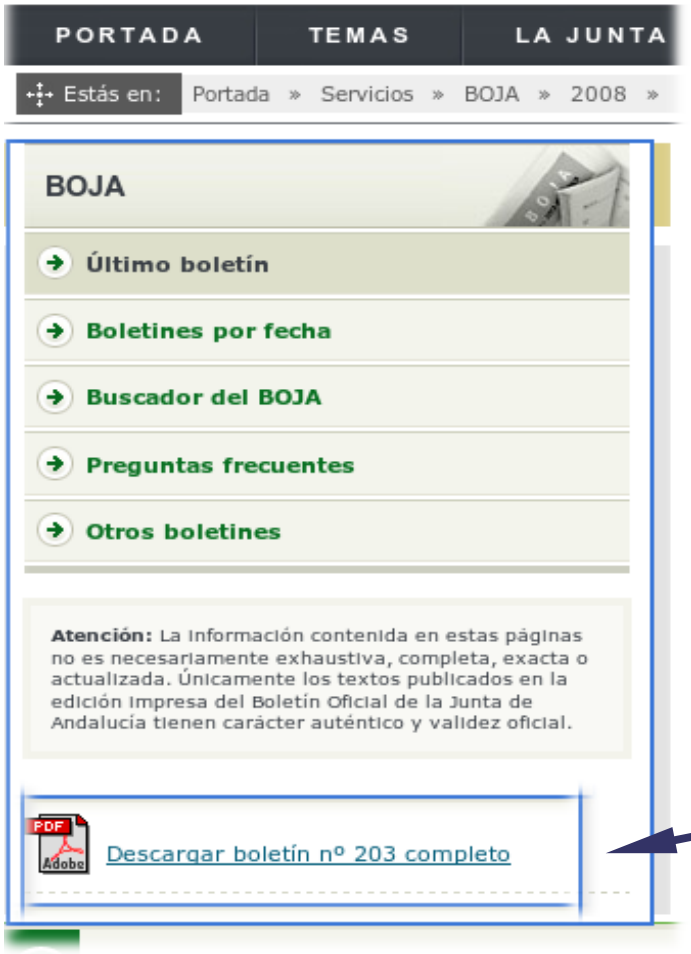
droids

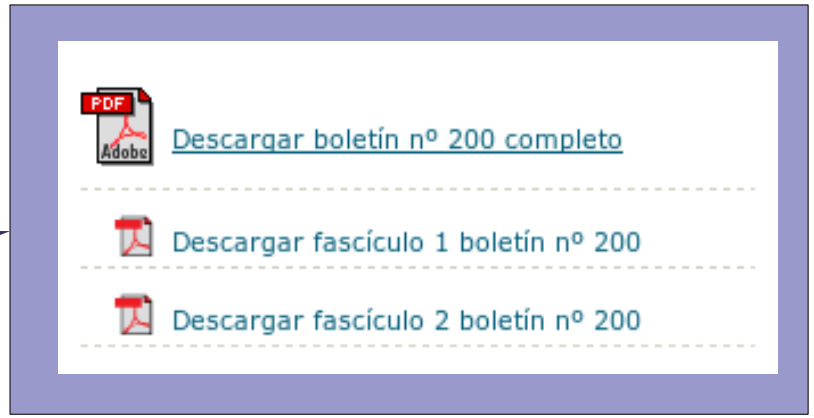
- bulk import of sources
  - crawl external site importing year ranges of official gazettes
- bulk task execution on repository (~500.000 dispositions)
  - update solr with mass content changes
  - change file properties (e.g. date format)
  - create fascicles descriptor
  - generate bulk html updates

httpd

The Apache HTTP Server Project is an effort to develop and maintain an open-source HTTP server for modern operating systems including UNIX and Windows NT. The goal of this project is to provide a secure, efficient and extensible server that provides HTTP services in sync with the current HTTP standards.

httpd





- Making static files semi dynamic
  - Split page in parts (each part served from a different html)

httpd ssi with forrest

- contract to inject ssi instruction

```
<forrest:content><forrest:part>
  <!--Set string-->
  <xsl:comment>#set var="map" value="/$REWRITEMAP_RESULT"</xsl:comment
  <!-- get year/number from map -->
  <xsl:comment>#if expr="$map == /^\/boletines\/(\\d{4})\/(\\d+)/" </xsl:comment>
  <xsl:comment>#set var="mapYear" value="$1" </xsl:comment>
  <xsl:comment>#set var="mapNumber" value="$2" </xsl:comment>
  <xsl:comment>#endif </xsl:comment>
  <!-- get year/number from request -->
  <xsl:comment>#if expr="$REQUEST_URI == /^\/boletines\/(\\d{4})\/(\\d+)/"
    </xsl:comment>
  <xsl:comment>#set var="year" value="$1" </xsl:comment>
  <xsl:comment>#set var="number" value="$2" </xsl:comment>
  <xsl:comment>#endif </xsl:comment>
  <!-- ... -->
```

httpd ssi with forrest

- contract to inject ssi instruction

```
<!-- ... -->
  <ul>
    <!-- compare both and set the focus -->
    <xsl:comment>#if expr="$year=$mapYear &amp;&amp;
      $number=$mapNumber" </xsl:comment>
    <li><a accesskey="U" class="actual" href="/BOJA">Último boletín</a></li>
    <li><a accesskey="F" href="/boja/boletines/">Boletines por fecha</a></li>
    <xsl:comment>#else</xsl:comment>
    <li><a accesskey="U" href="/BOJA">Último boletín</a></li>
    <li><a accesskey="F" class="actual" href="/boja/boletines/"> Boletines por
      fecha</a></li>
    <xsl:comment>#endif</xsl:comment> ... </ul> ...
  <div id="texto_informativo"><p><strong>Atención:</strong> ...</p></div>
 </forrest:part>
</forrest:content>
```

httpd

- Making static files semi dynamic
  - Page created with ssi injection contract mentioned before
  - Activate the rewrite

```
RewriteMap portadaboja txt:/opt/datos/httpd/redirect.txt

RewriteRule ^(.*)     %{DOCUMENT_ROOT}$1
   [E=REWRITEMAP_RESULT: ${portadaboja:boletin},L]
```

next steps

- Using JCR repository to store internal data (Sling/Jackrabbit)
- Replace Tomcat with Felix
- Creating admin interface for solr
- Creating web admin interface for droids

ApacheCon

Leading the Wave of Open Source

ApacheCon

Thank you for your attention

Thorsten Scherler
Sociedad Andaluza para el Desarrollo
de la Sociedad de la Información S.A.U. (SADESI)

thorsten@apache.org
ApacheCon US 2008, New Orleans, 07 November 2008